Exploratory Data Analysis

Lecture 3

Corina Besliu

Technical University of Moldova

September 10, 2021

Outline

- Types of Data
- EDA Methods by Type of Data
- Graphical Methods Nominal Data
- Graphical Methods for Qunatitative (Interval) Data

Outline

- Types of Data
- EDA Methods by Type of Data
- Graphical Methods Nominal Data
- Graphical Methods for Qunatitative (Interval) Data

Outline

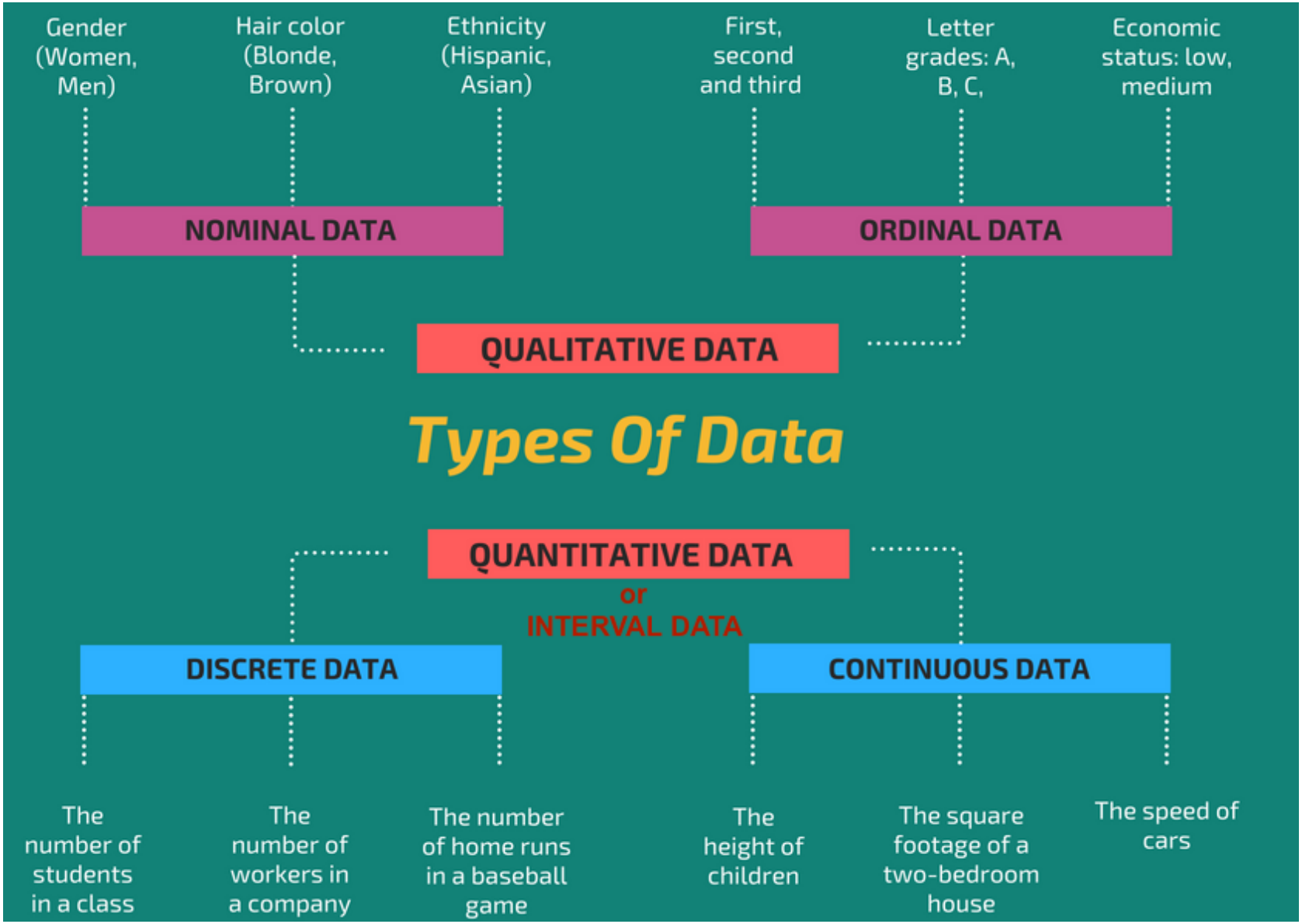
- Types of Data
- EDA Methods by Type of Data
- Graphical Methods Nominal Data
- Graphical Methods for Qunatitative (Interval) Data

Outline

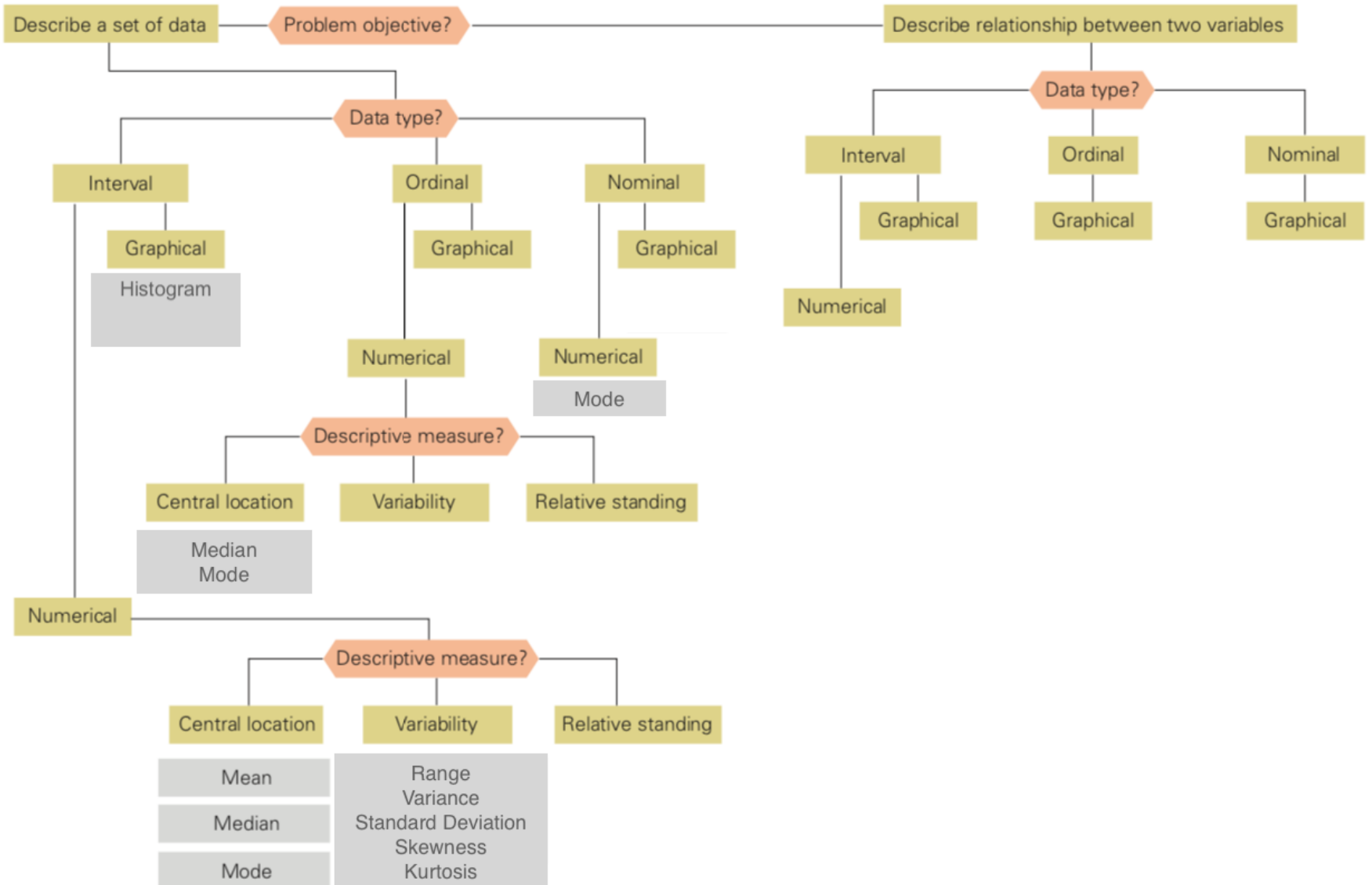
- Types of Data
- EDA Methods by Type of Data
- Graphical Methods Nominal Data
- Graphical Methods for Qunatitative (Interval) Data

Lecture Outline

Types of Data



EDA Methods by Type of Data



Lecture Outline

Work Status General Social Survey 2020

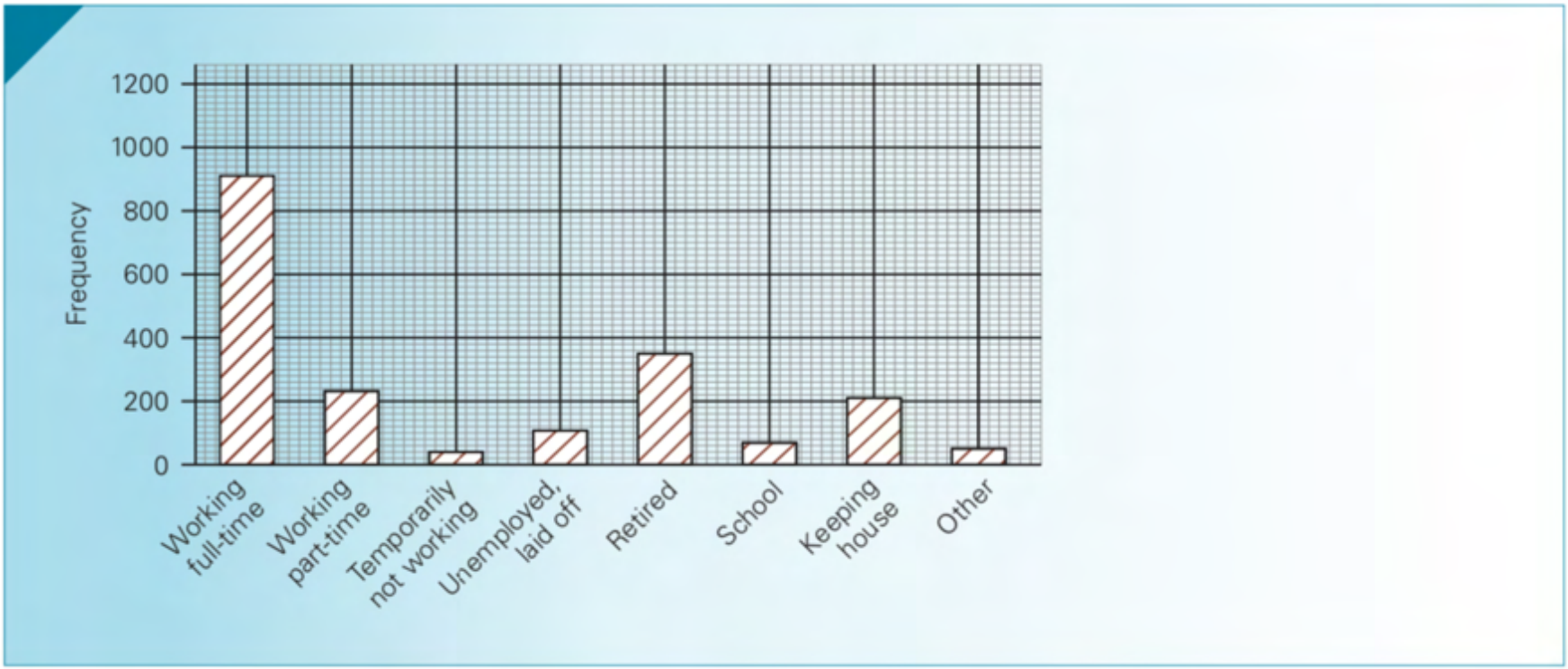
The only allowable calculation on nominal data is to count the frequency or compute the percentage that each value of the variable represents.

Frequency Distribution and Relative Frequency Distribution

WORK STATUS	CODE	FREQUENCY	RELATIVE FREQUENCY (%)
Working full-time	1	912	46.2
Working part-time	2	226	11.5
Temporarily not working	3	40	2.0
Unemployed, laid off	4	104	5.3
Retired	5	357	18.1
School	6	70	3.5
Keeping house	7	210	10.6
Other	8	54	2.7
Total		1973	100

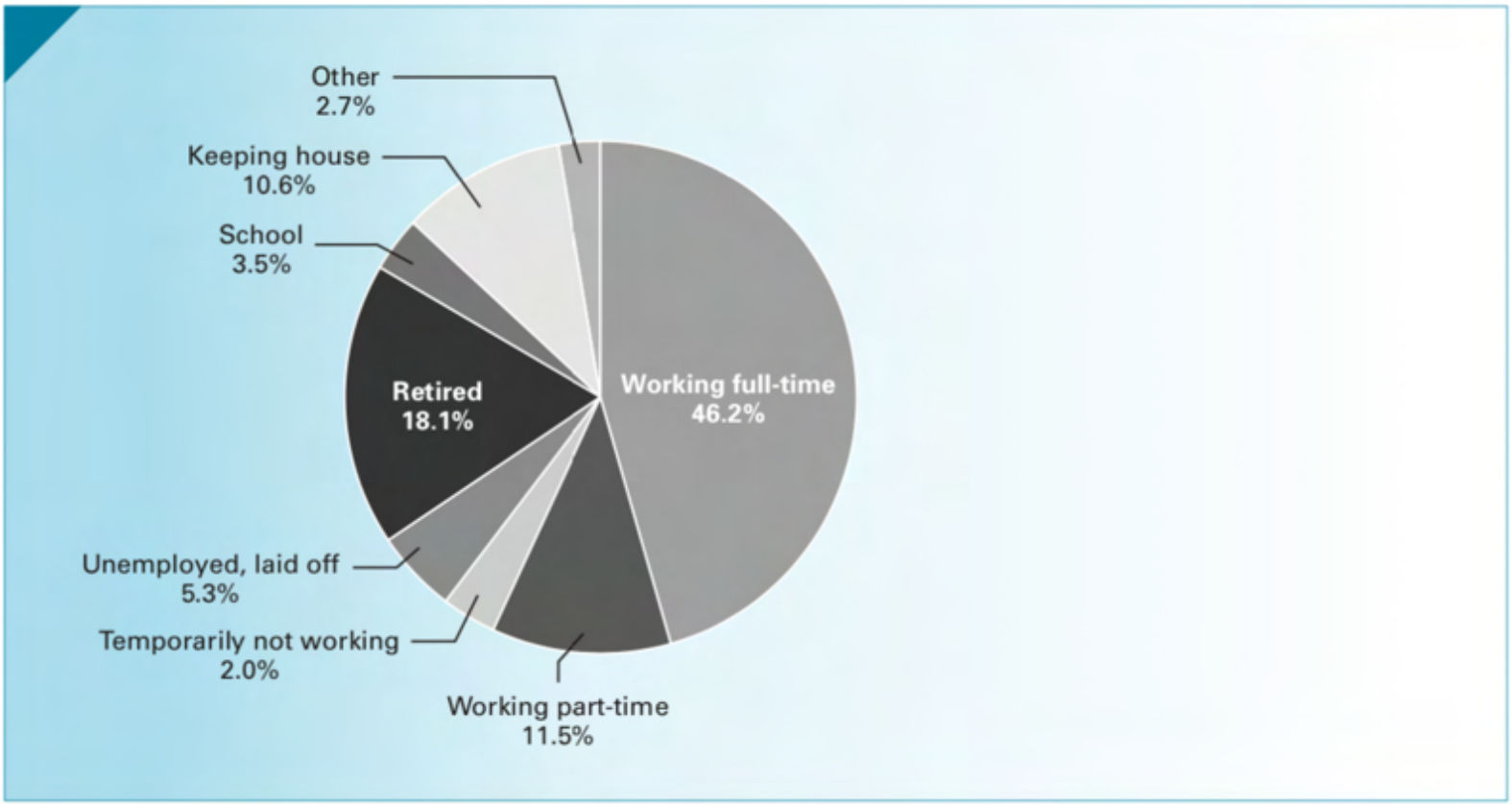
Bar Chart

To visualize the Frequency Distribution use Bar Charts



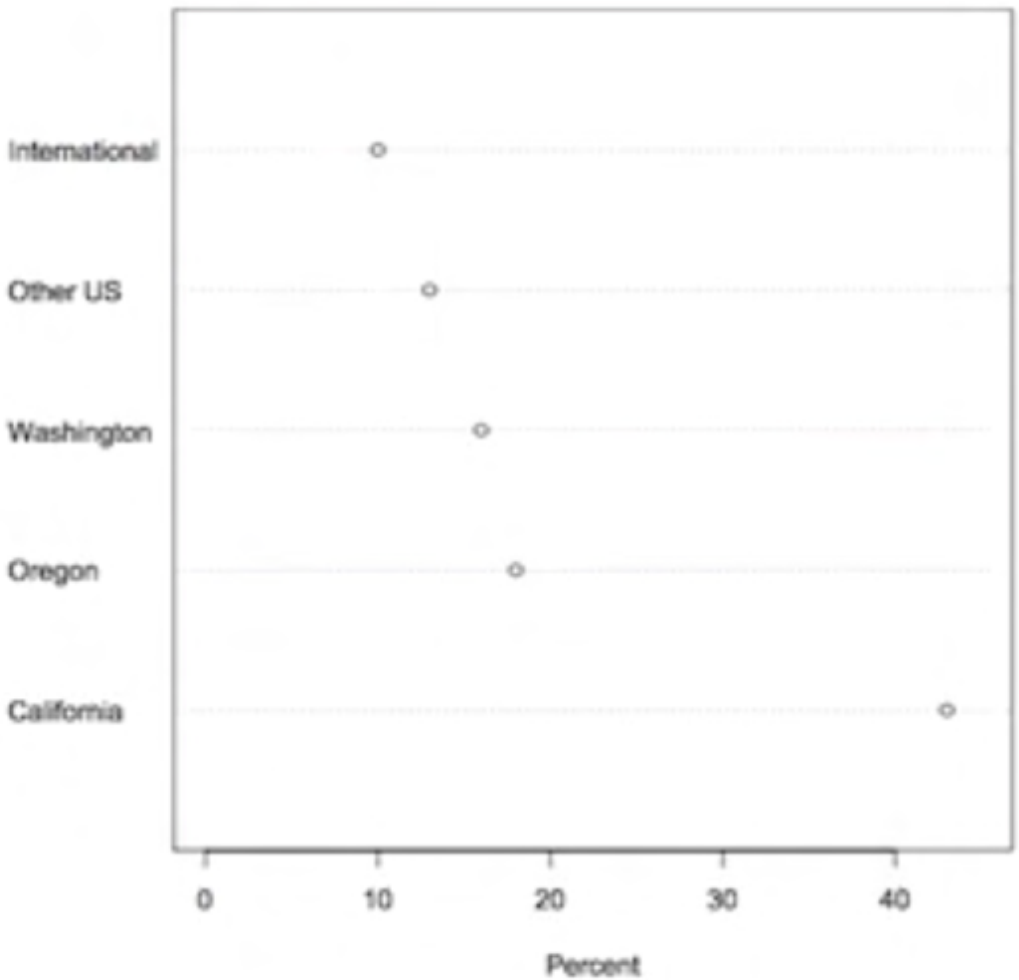
Pie Chart

To emphasize the relative frequencies use Pie Charts



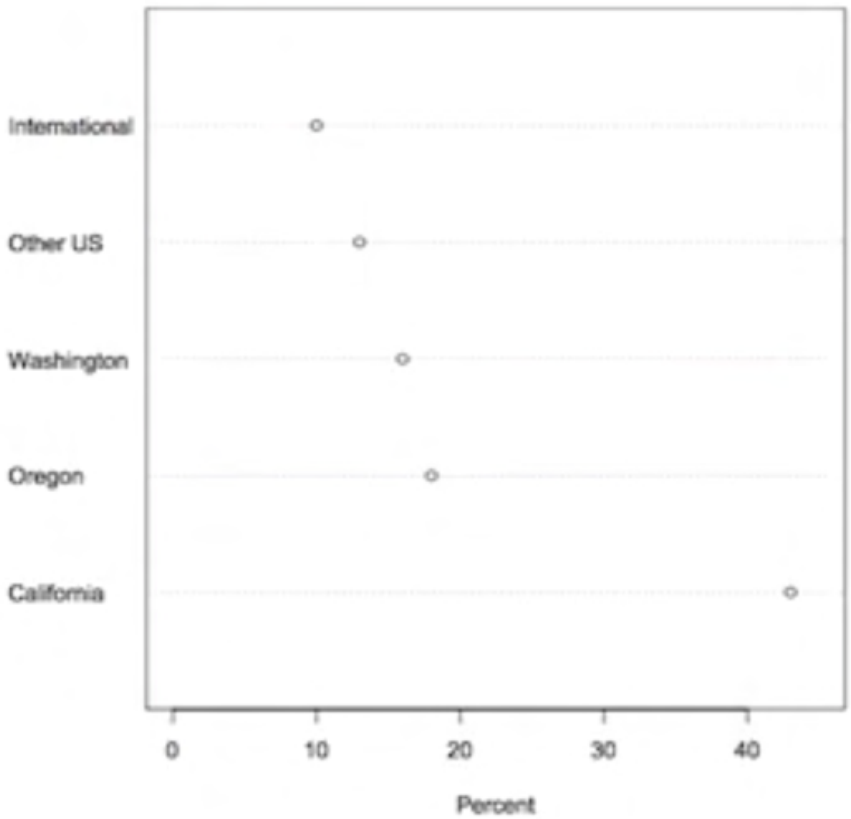
Dot Plot

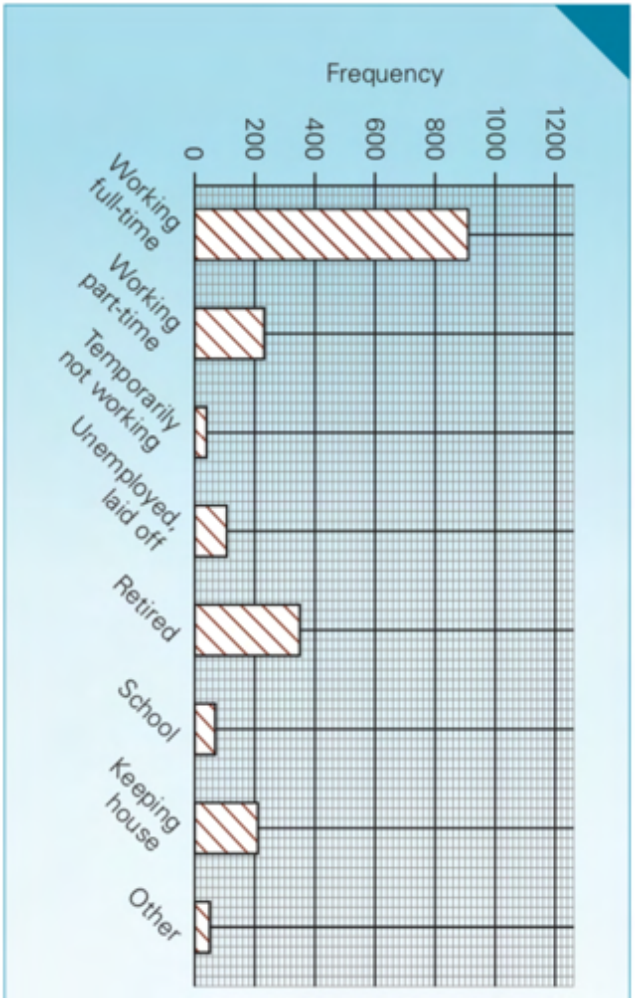
To compare frequencies of different categories use Dot Plots.



Dot Plot

A Dot Plot is esentially a Bar Chart put on it's side





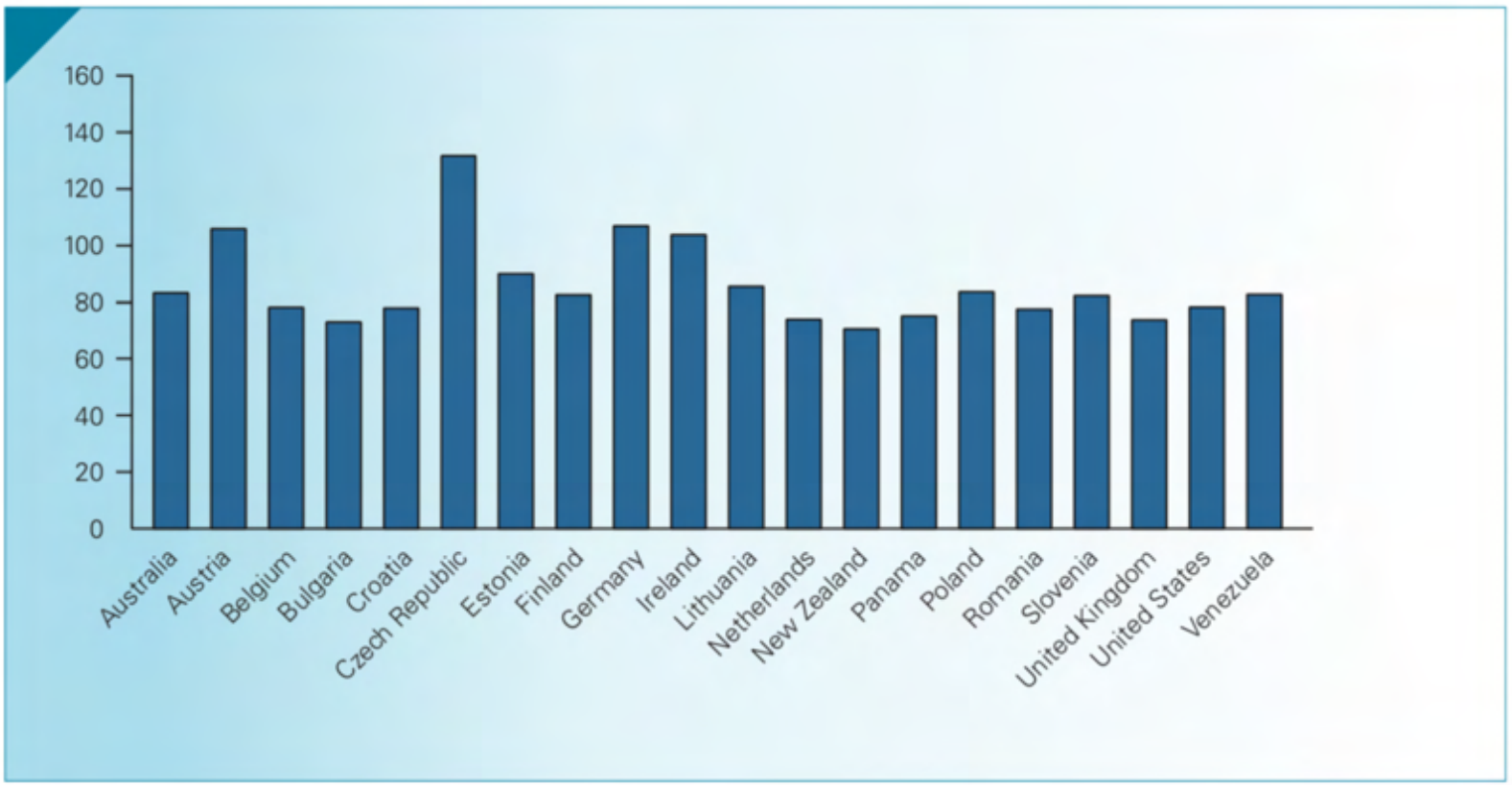
Bier Consumption by Country

Treat Country as the Category and Beer Consumption as Frequency

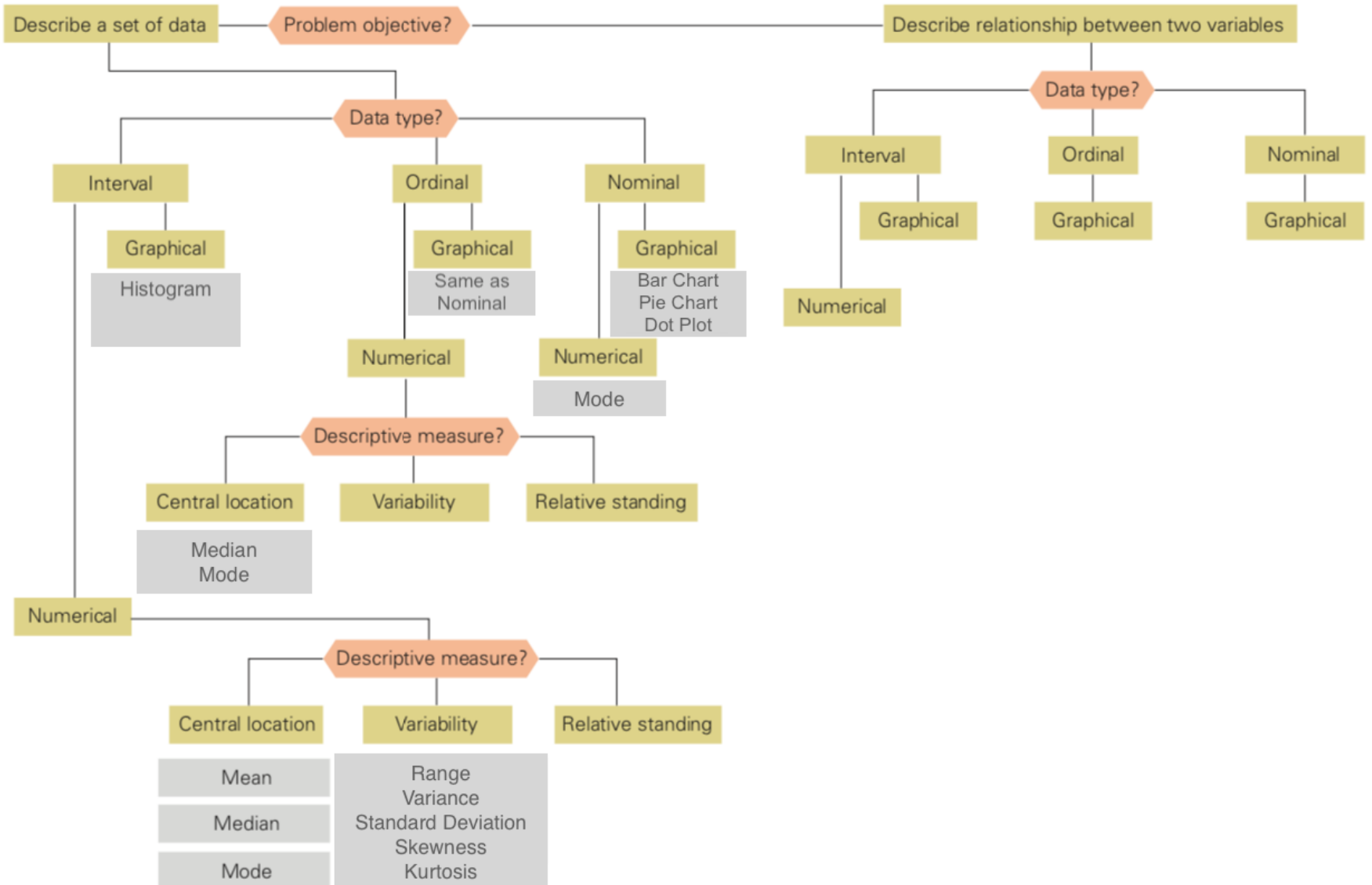
COUNTRY	BEER CONSUMPTION (L/YR)
Australia	83.4
Austria	105.8
Belgium	78.0
Bulgaria	72.8
Croatia	77.8
Czech Republic	131.7
Estonia	90.6
Finland	82.7
Germany	106.8
Ireland	103.7
Lithuania	85.7
Netherlands	73.9
New Zealand	70.5
Panama	75.0
Poland	83.6
Romania	77.4
Slovenia	82.7
United Kingdom	73.7
United States	78.2
Venezuela	83.0

Bier Consumption by Country Bar Chart

Countries (i.e. categories) = X-axis
Consumption (i.e. frequencies) = Y-axis



EDA Graphical Methods (Nominal & Ordinal Data)



Lecture Outline

Telephone Company's Problem

- Telephone companies competed on price because the service each offers is similar.
- Pricing a service or product in the face of stiff competition is very difficult.
- One must understand very well customer behavior!
- The general manager of a phone company conducts a research of 200 new subscribers and records their first months bills.
- He planned to present his findings to senior executives.

Telephone Company's Problem

- Telephone companies competed on price because the service each offers is similar.
- Pricing a service or product in the face of stiff competition is very difficult.
- One must understand very well customer behavior!
- The general manager of a phone company conducts a research of 200 new subscribers and records their first months bills.
- He planned to present his findings to senior executives.

Telephone Company's Problem

- Telephone companies competed on price because the service each offers is similar.
- Pricing a service or product in the face of stiff competition is very difficult.
- One must understand very well customer behavior!
- The general manager of a phone company conducts a research of 200 new subscribers and records their first months bills.
- He planned to present his findings to senior executives.

Telephone Company's Problem

- Telephone companies competed on price because the service each offers is similar.
- Pricing a service or product in the face of stiff competition is very difficult.
- One must understand very well customer behavior!
- The general manager of a phone company conducts a research of 200 new subscribers and records their first months bills.
- He planned to present his findings to senior executives.

Telephone Company's Problem

- Telephone companies competed on price because the service each offers is similar.
- Pricing a service or product in the face of stiff competition is very difficult.
- One must understand very well customer behavior!
- The general manager of a phone company conducts a research of 200 new subscribers and records their first months bills.
- He planned to present his findings to senior executives.

200 Telephone Bills

First Month Telephone Bills New Customers

42.19	39.21	75.71	8.37	1.62	28.77	35.32	13.9	114.67	15.3
38.45	48.54	88.62	7.18	91.1	9.12	117.69	9.22	27.57	75.49
29.23	93.31	99.5	11.07	10.88	118.75	106.84	109.94	64.78	68.69
89.35	104.88	85	1.47	30.62	0	8.4	10.7	45.81	35
118.04	30.61	0	26.4	100.05	13.95	90.04	0	56.04	9.12
110.46	22.57	8.41	13.26	26.97	14.34	3.85	11.27	20.39	18.49
0	63.7	70.48	21.13	15.43	79.52	91.56	72.02	31.77	84.12
72.88	104.84	92.88	95.03	29.25	2.72	10.13	7.74	94.67	13.68
83.05	6.45	3.2	29.04	1.88	9.63	5.72	5.04	44.32	20.84
95.73	16.47	115.5	5.42	16.44	21.34	33.69	33.4	3.69	100.04
103.15	89.5	2.42	77.21	109.08	104.4	115.78	6.95	19.34	112.94
94.52	13.36	1.08	72.47	2.45	2.88	0.98	6.48	13.54	20.12
26.84	44.16	76.69	0	21.97	65.9	19.45	11.64	18.89	53.21
93.93	92.97	13.62	5.64	17.12	20.55	0	83.26	1.57	15.3
90.26	99.56	88.51	6.48	19.7	3.43	27.21	15.42	0	49.24
72.78	92.62	55.99	6.95	6.93	10.44	89.27	24.49	5.2	9.44
101.36	78.89	12.24	19.6	10.05	21.36	14.49	89.13	2.8	2.67
104.8	87.71	119.63	8.11	99.03	24.42	92.17	111.14	5.1	4.69
74.01	93.57	23.31	9.01	29.24	95.52	21	92.64	3.03	41.38
56.01	0	11.05	84.77	15.21	6.72	106.59	53.9	9.16	45.77

But what information can be extracted from these data?

200 Telephone Bills

First Month Telephone Bills New Customers

42.19	39.21	75.71	8.37	1.62	28.77	35.32	13.9	114.67	15.3
38.45	48.54	88.62	7.18	91.1	9.12	117.69	9.22	27.57	75.49
29.23	93.31	99.5	11.07	10.88	118.75	106.84	109.94	64.78	68.69
89.35	104.88	85	1.47	30.62	0	8.4	10.7	45.81	35
118.04	30.61	0	26.4	100.05	13.95	90.04	0	56.04	9.12
110.46	22.57	8.41	13.26	26.97	14.34	3.85	11.27	20.39	18.49
0	63.7	70.48	21.13	15.43	79.52	91.56	72.02	31.77	84.12
72.88	104.84	92.88	95.03	29.25	2.72	10.13	7.74	94.67	13.68
83.05	6.45	3.2	29.04	1.88	9.63	5.72	5.04	44.32	20.84
95.73	16.47	115.5	5.42	16.44	21.34	33.69	33.4	3.69	100.04
103.15	89.5	2.42	77.21	109.08	104.4	115.78	6.95	19.34	112.94
94.52	13.36	1.08	72.47	2.45	2.88	0.98	6.48	13.54	20.12
26.84	44.16	76.69	0	21.97	65.9	19.45	11.64	18.89	53.21
93.93	92.97	13.62	5.64	17.12	20.55	0	83.26	1.57	15.3
90.26	99.56	88.51	6.48	19.7	3.43	27.21	15.42	0	49.24
72.78	92.62	55.99	6.95	6.93	10.44	89.27	24.49	5.2	9.44
101.36	78.89	12.24	19.6	10.05	21.36	14.49	89.13	2.8	2.67
104.8	87.71	119.63	8.11	99.03	24.42	92.17	111.14	5.1	4.69
74.01	93.57	23.31	9.01	29.24	95.52	21	92.64	3.03	41.38
56.01	0	11.05	84.77	15.21	6.72	106.59	53.9	9.16	45.77

But what information can be extracted from these data?

Possible Questions

Executives might be interested to know:

- How the numbers are distributed between the minimul bill of 0 and the maximum of 119.63.
- Are there many small bills and few large bills?
- What is the typical bill?
- Are the bills somewhat similar or do they vary considerably?

Solution:

Construct a frequency distribution from which a histogram can be drawn.

Possible Questions

Executives might be interested to know:

- How the numbers are distributed between the minimul bill of 0 and the maximum of 119.63.
- Are there many small bills and few large bills?
- What is the typical bill?
- Are the bills somewhat similar or do they vary considerably?

Solution:

Construct a frequency distribution from which a histogram can be drawn.

Possible Questions

Executives might be interested to know:

- How the numbers are distributed between the minimul bill of 0 and the maximum of 119.63.
- Are there many small bills and few large bills?
- What is the typical bill?
- Are the bills somewhat similar or do they vary considerably?

Solution:

Construct a frequency distribution from which a histogram can be drawn.

Possible Questions

Executives might be interested to know:

- How the numbers are distributed between the minimul bill of 0 and the maximum of 119.63.
- Are there many small bills and few large bills?
- What is the typical bill?
- Are the bills somewhat similar or do they vary considerably?

Solution:

Construct a frequency distribution from which a histogram can be drawn.

Possible Questions

Executives might be interested to know:

- How the numbers are distributed between the minimul bill of 0 and the maximum of 119.63.
- Are there many small bills and few large bills?
- What is the typical bill?
- Are the bills somewhat similar or do they vary considerably?

Solution:

Construct a frequency distribution from which a histogram can be drawn.

Choosing the Intervals

The manager has chosen eight classes defined in such a way that each observation falls into one and only one class.

CLASS LIMITS	FREQUENCY
0 to 15*	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200

- The intervals do not overlap, so there is no uncertainty about which interval to assign to any observation.
- The intervals are equally wide.

Choosing the Intervals

The manager has chosen eight classes defined in such a way that each observation falls into one and only one class.

CLASS LIMITS	FREQUENCY
0 to 15*	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200

- The intervals do not overlap, so there is no uncertainty about which interval to assign to any observation.
- The intervals are equally wide.

Choosing the Intervals

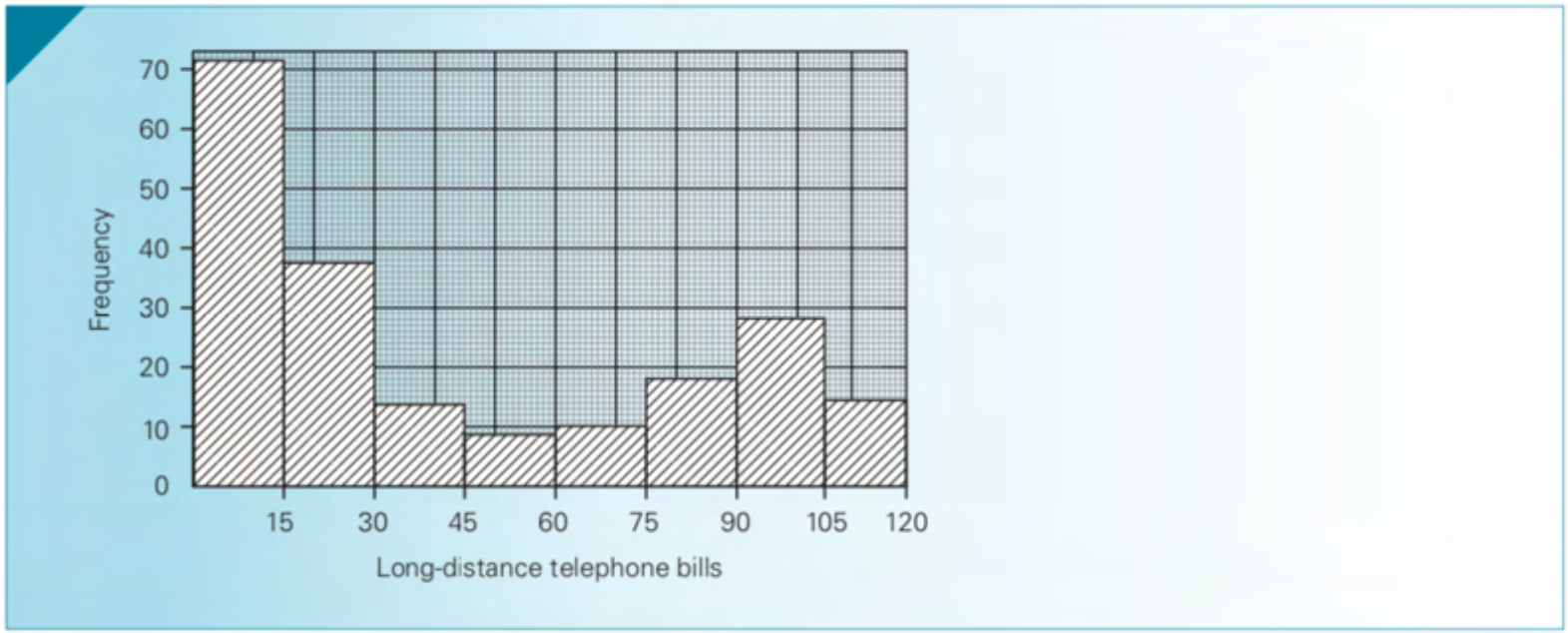
The manager has chosen eight classes defined in such a way that each observation falls into one and only one class.

CLASS LIMITS	FREQUENCY
0 to 15*	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200

- The intervals do not overlap, so there is no uncertainty about which interval to assign to any observation.
- The intervals are equally wide.

Telephone Bills Histogram

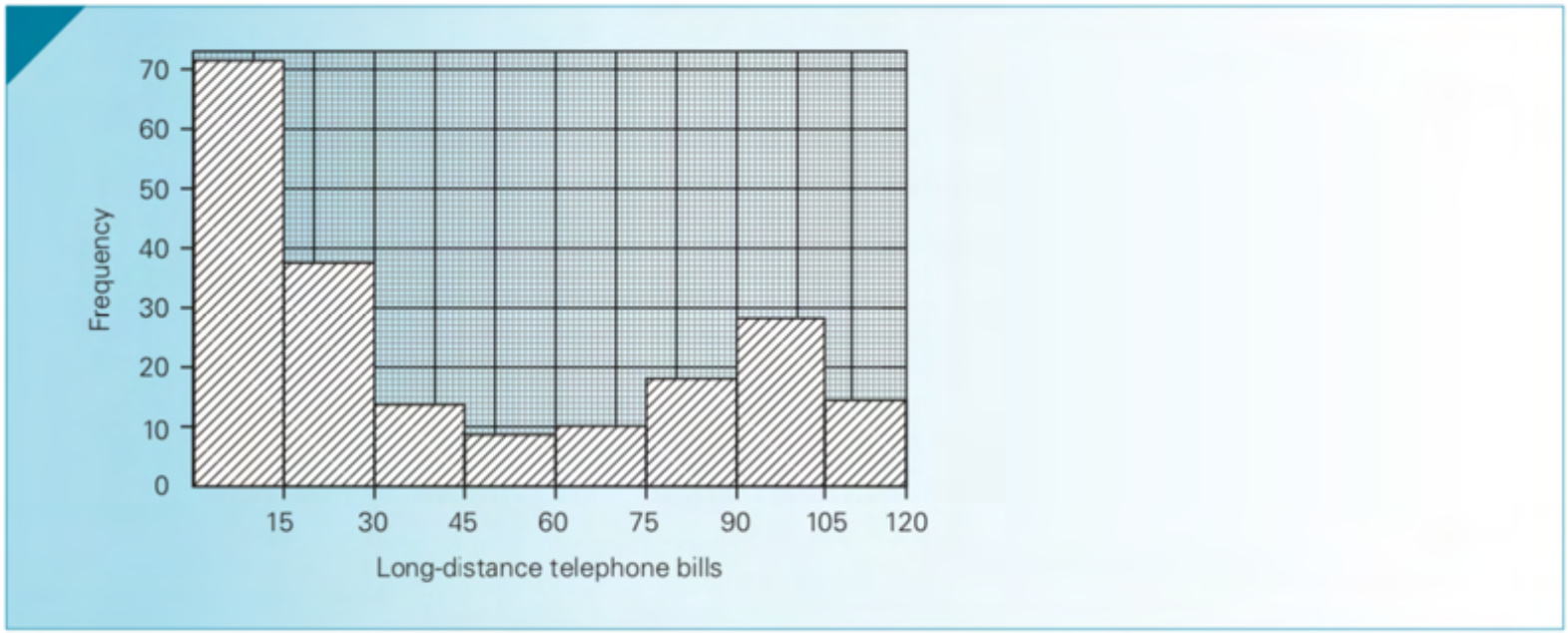
A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies.



- About half the monthly bills are small ($0 to $30)
- A few bills are in the middle range ($30 to $75)
- A relatively large number of long-distance bills are at the high end of the range.

Telephone Bills Histogram

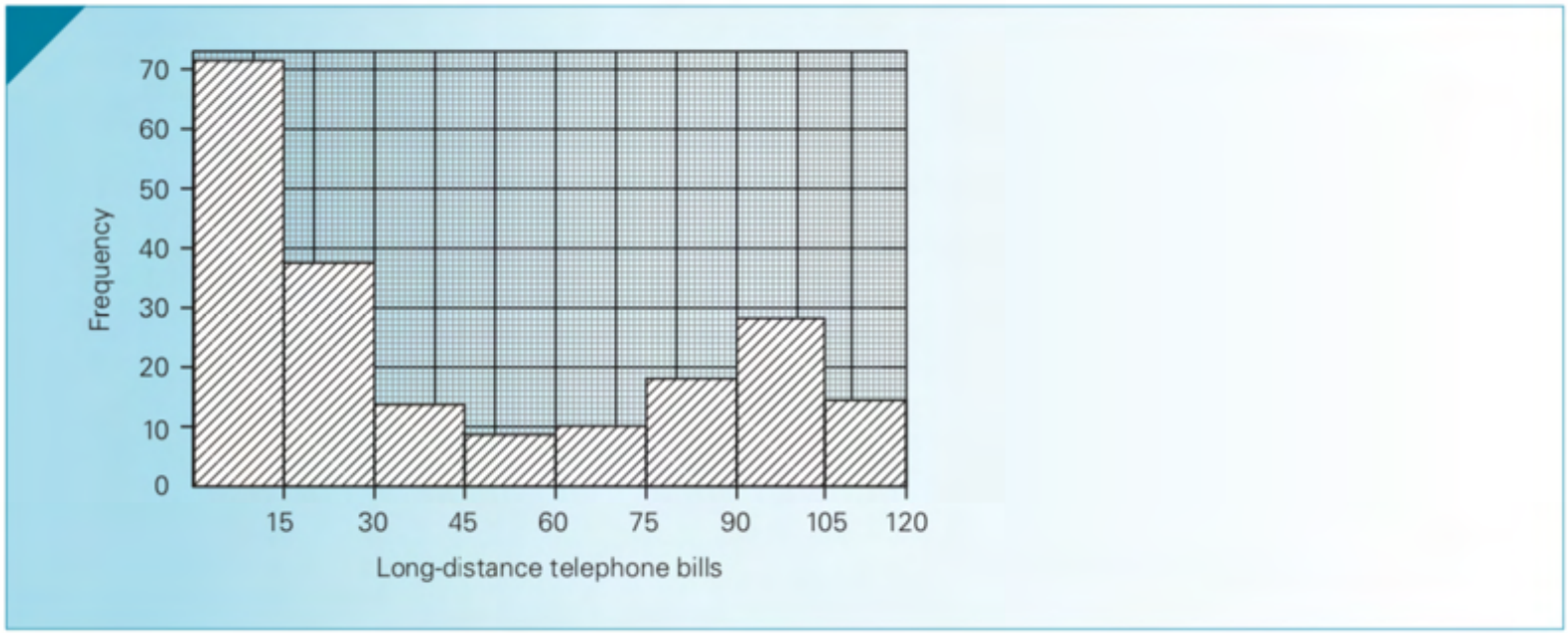
A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies.



- About half the monthly bills are small ($0 to $30)
- A few bills are in the middle range ($30 to $75)
- A relatively large number of long-distance bills are at the high end of the range.

Telephone Bills Histogram

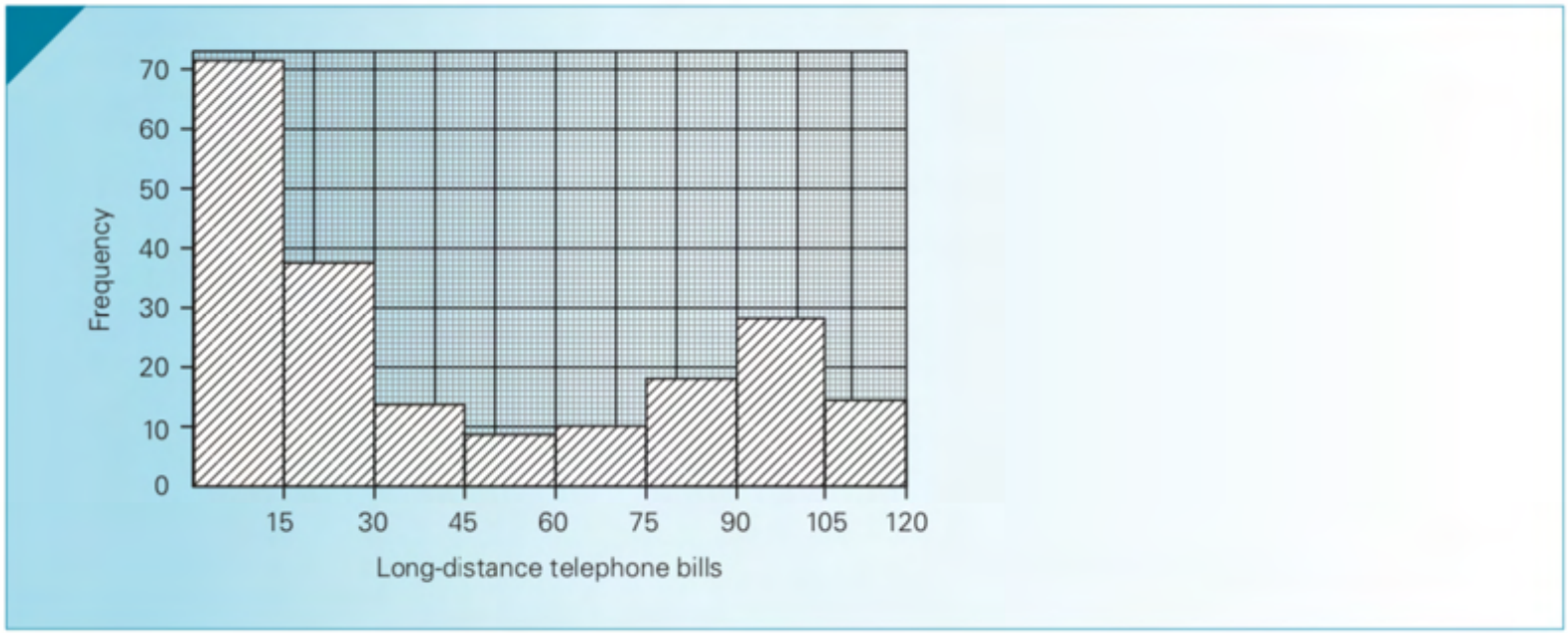
A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies.



- About half the monthly bills are small ($0 to $30)
- A few bills are in the middle range ($30 to $75)
- A relatively large number of long-distance bills are at the high end of the range.

Telephone Bills Histogram

A histogram is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies.



- About half the monthly bills are small ($0 to $30)
- A few bills are in the middle range ($30 to $75)
- A relatively large number of long-distance bills are at the high end of the range.

Selecting Nb. of Class Intervals

The nb. of class intervals depends entirely on the nb. of observations.

NUMBER OF OBSERVATIONS	NUMBER OF CLASSES
Less than 50	5–7
50–200	7–9
200–500	9–10
500–1,000	10–11
1,000–5,000	11–13
5,000–50,000	13–17
More than 50,000	17–20

- An alternative to the guidelines is to use Sturgess formula:

$$\text{Number of class intervals} = 1 + 3.3 \log(n)$$

- Eg: $n = 50 \Rightarrow$ Number of class intervals $= 1 + 3.3 \log(50) = 6$.

Selecting Nb. of Class Intervals

The nb. of class intervals depends entirely on the nb. of observations.

NUMBER OF OBSERVATIONS	NUMBER OF CLASSES
Less than 50	5–7
50–200	7–9
200–500	9–10
500–1,000	10–11
1,000–5,000	11–13
5,000–50,000	13–17
More than 50,000	17–20

- An alternative to the guidelines is to use Sturgess formula:

$$\text{Number of class intervals} = 1 + 3.3 \log (n)$$

- Eg: $n = 50 \Rightarrow$ Number of class intervals $= 1 + 3.3 \log (50) = 6$.

Selecting Nb. of Class Intervals

The nb. of class intervals depends entirely on the nb. of observations.

NUMBER OF OBSERVATIONS	NUMBER OF CLASSES
Less than 50	5–7
50–200	7–9
200–500	9–10
500–1,000	10–11
1,000–5,000	11–13
5,000–50,000	13–17
More than 50,000	17–20

- An alternative to the guidelines is to use Sturgess formula:

$$\text{Number of class intervals} = 1 + 3.3 \log (n)$$

- Eg: $n = 50 \Rightarrow$ Number of class intervals $= 1 + 3.3 \log (50) = 6$.

Types of Histograms

FIGURE **3.2** Three Symmetric Histograms

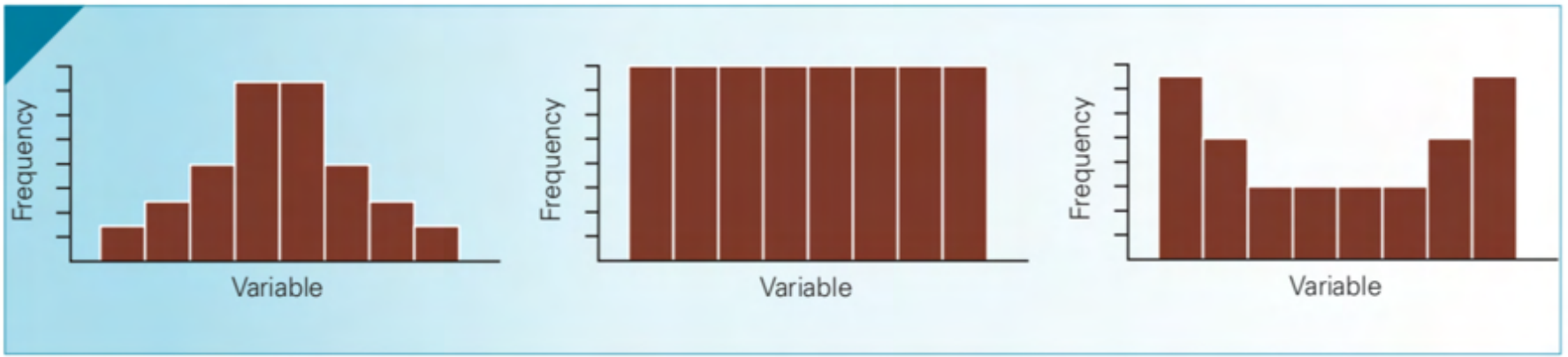
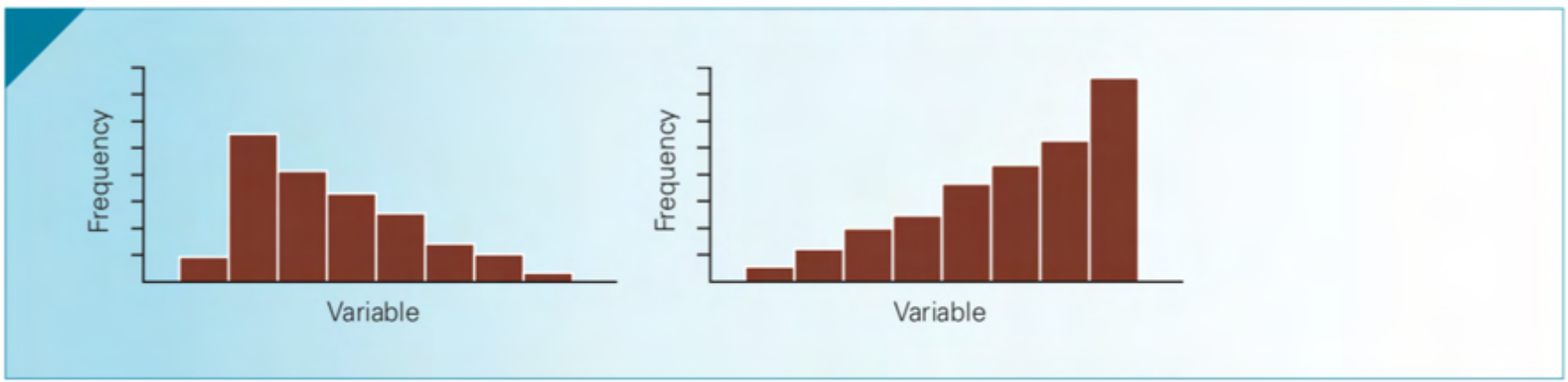


FIGURE **3.3** Positively and Negatively Skewed Histograms



Types of Histograms

FIGURE **3.4** A Unimodal Histogram

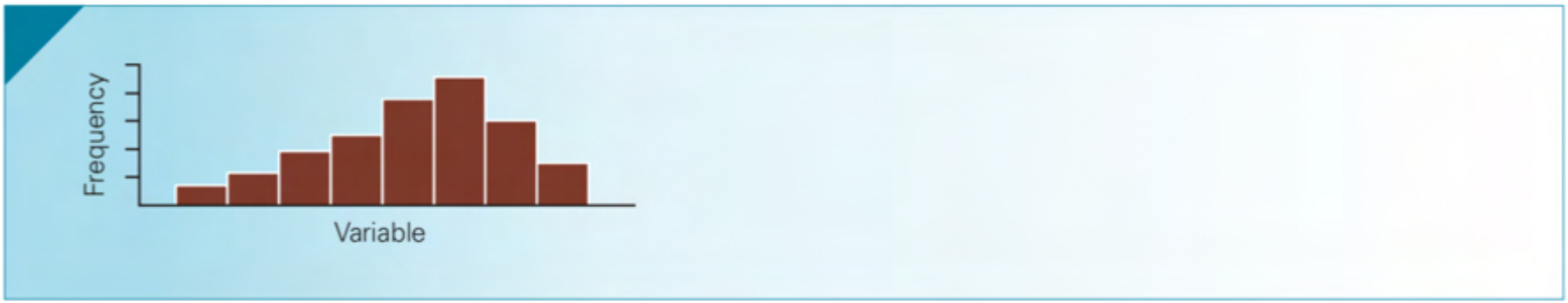
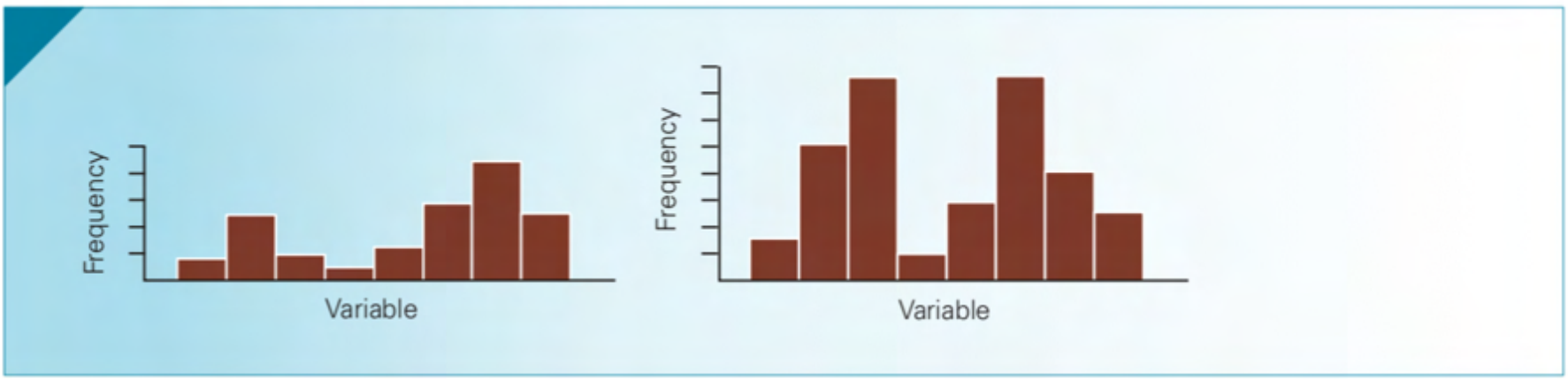
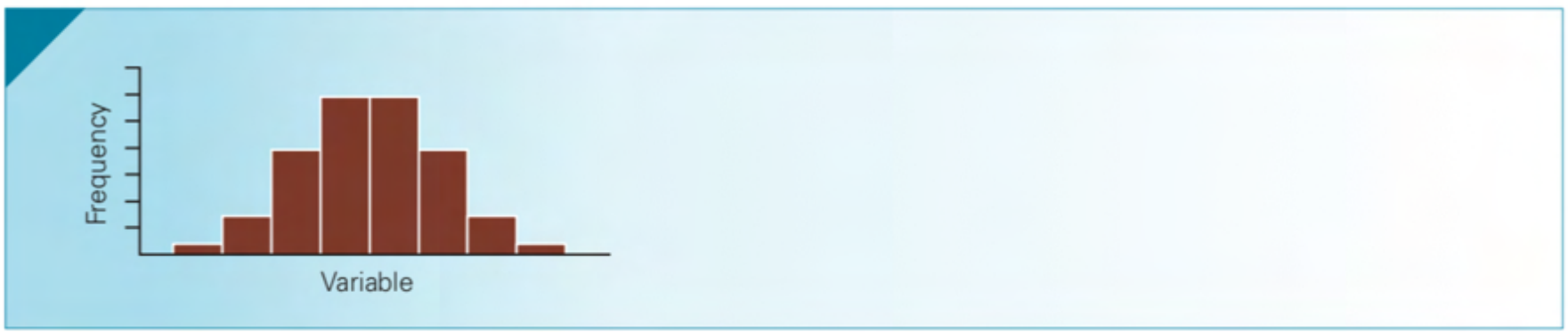


FIGURE **3.5** Bimodal Histograms



Types of Histograms

FIGURE **3.6** **Bell-Shaped Histogram**



Stem-and-Leaf Display

- With histograms we lose potentially useful information by classifying the observations.
- Split each observation into two parts, a stem and a leaf.

Stem	Leaf
0	000000001111122222233333455555566666667788889999999
1	000001111233333334455555667889999
2	0000111112344666778999
3	001335589
4	124445589
5	33566
6	3458
7	022224556789
8	334457889999
9	00112222233344555999
10	001344446699
11	0124557889

Lecture Outline

Let's get Started!

Access Google Colaboratory through your Gmail account